HARMONY SEARCH-BASED FUZZY CLUSTERING ALGORITHMS FOR IMAGE SEGMENTATION

OSAMA MOH'D RADI ALIA

UNIVERSITI SAINS MALAYSIA

2011

HARMONY SEARCH-BASED FUZZY CLUSTERING ALGORITHMS FOR IMAGE SEGMENTATION

by

OSAMA MOH'D RADI ALIA

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

February 2011

ACKNOWLEDGEMENTS

IN THE NAME OF ALLAH THE ALL-COMPASSIONATE, ALL-MERCIFIFUL

All thanks are due to ALLAH, the most merciful, the most compassionate. For the countless blessings He has bestowed upon me.

Then, I would also like to express my deepest gratitude towards my thesis advisor, Associate Professor Dr. Mandava Rajeswari, for her continuous wisdom, encouragement, friendship, guidance and support over this course of my study. I am forever grateful for the chance to work in her research group.

Thank you also to Dr. Dhanesh Ramachandram for all the useful ideas he has given. His assistance and useful advice in publications-related issues are highly appreciated.

I would like to thank my co-supervisors, Associate Professor Dr. Mohd Ezane Aziz and Dr. Nor Azman Mat Zain of USM Hospital for spending much of their valuable time explaining the medical terminologies and issues related to malignant bone tumour “osteosarcoma”. These have been great sources of knowledge for guiding this work, especially in the automatic segmentation and labeling of different types of tumour cells. I also highly appreciate the comments and the many suggestions regarding the quality of this work, as well as how to make the developed tools more applicable in the setting of medical research and clinical use.

I would like to thank Prof. Ibrahim Lutifi Shuaib for his help and useful discussions, without which this thesis would not have been able to reach its completion. Not forgetting friends and labmates at the Computer Vision Research Group at Universiti Sains Malaysia, who have been a tremendous source of support and a lot of fun to work with. I will forever remember moments together. Thank you very-very much Mahmoud, Ahmad, Alfian, Mozaher, Ehsan, Idayu, and Anusha. Special thanks also to Ali Alghafri, Khalid Jaber, Mohammad Said Abu Al-Rub, Asef Al-khatib and Mohammad Al-Betar for their support.

Special thanks to School of Computer Sciences USM, for providing a conducive environment during the course of my research. The researcher is also deeply grateful to the Research Assistance Scheme at Universiti Sains Malaysia, USM Fellowship Scheme provided by the Institute of Graduate Studies (IPS), USM, and "Universiti Sains Malaysia Research University Grant" titled "Delineation and visualization of Tumour and Risk Structures - DVTRS" under grant number $1001/PKOMP/817001$, which supported my living cost during my study and research at the campus. I would like to thank the USM members for giving me the chance to gain my degree, and for providing me and the other students excellent facilities to conduct research.

Last but not least, may Allah have great compassion towards my parents' souls. Gratitude beyond words to my beloved wife Noor; lovely sons Abdalrahman and Zain Alabideen; beautiful daughter Maria; brothers, Radi, Khaldon, and Adnan; and sister, Ameera, who have given constant love, support, encouragement, and patience.

Thank you all

LIST OF PUBLICATIONS

Journal Publications:

1. Alia O. M. and Mandava R. and Aziz M. E. (2011). A Hybrid Harmony Search Algorithm for MRI Brain Segmentation, DOI: 10.1007/s12065-011-0048-1, *Evolutionary Intelligence, Springer.*

2. Alia O. M. and Mandava R. (2011). The Variants of the Harmony Search Algorithm - An Overview , DOI: 10.1007/s10462-010-9201-y, *Artificial Intelligence Review, Springer.*

Conference Publications:

1. Alia O. M. and Mandava R. (2011). Data Clustering using Harmony Search Algorithm, *The 15th Online Conference on Soft Computing in Industrial Applications, (WSC15), 15-27 November 2010, Advances in Intelligent and Soft Computing, Springer Berlin/ Heidelberg.*

2. Mandava R. and Alia O. M. and Wei B. C. and Ramachandram D. and Aziz M. E. and Shuaib I. L. (2010). Osteosarcoma Segmentation in MRI using Dynamic Harmony Search Based Clustering, *International Conference of Soft Computing and Pattern Recognition, SOCPAR10, 7-10/12/2010, Cergy-Pontoise, France.*

3. Alia O. M. and Mandava R. and Aziz M. E. (2010). A Hybrid Harmony Search Algorithm to MRI Brain Segmentation, *The 9th IEEE International Conference on Cognitive Informatics, ICCI2010, pp. 712-719, Beijing, China.*

4. Alia O. M. and Mandava R. and Ramachandram D. and Aziz M. E. (2009). Dynamic fuzzy clustering using Harmony Search with application to image segmentation, *IEEE International Symposium on Signal Processing and Information Tech-*

nology, ISSPIT09, pp. 538-543, Ajman, UAE.

5. Alia O. M. and Mandava R. and Ramachandram D. and Aziz M. E. (2009). A Novel Image Segmentation Algorithm Based on Harmony Fuzzy Search Algorithm, *International Conference of Soft Computing and Pattern Recognition, SOCPAR09, pp. 335-340, Malacca, Malaysia.*

6. Alia O. M. and Mandava R. and Ramachandram D. and Aziz M. E. (2009). Harmony search-based cluster initialization for fuzzy c-means segmentation of MR images, *International Technical Conference of IEEE Region 10, TENCON, pp. 1-6, Singapore.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	Two-Dimensional
3D	Three-Dimensional
ABC	Artificial Bee Colony
ANN	Artificial Neural Network
ACO	Ant Colony Optimization
bw	bandwidth parameter
CSF	CerebroSpinal Fluid
CT	Computed Tomography
DCHS	Dynamic Fuzzy Clustering using the Harmony Search
DE	Differential Evolution
DICOM	Digital Imaging and Communications in Medicine
EA	Evolutionary Algorithm
EM	Expectation Maximization Algorithm
EOR	Empty Operator Rate
EP	Evolutionary Programming
ES	Evolutionary Strategies
FLAIR	Fluid Attenuation Inversion Recovery
FCM	Fuzzy C-Means Algorithm

FCMR Fuzzy C-Means Rate

GA Genetic Algorithm

GM Gray Matter

HFCM Harmony Fuzzy C-Means Algorithm

HFISA Harmony Fuzzy Image Segmentation Algorithm

HM Harmony Memory

HMCR Harmony Memory Considering Rate

HMS Harmony Memory Size

HS Harmony Search Algorithm

MRI Magnetic Resonance Imaging

MSLs Multiple Sclerosis Lesions

NI Number of Iterations

NP Non-deterministic Polynomial-time

PAR Pitch Adjusting Rate

PC Partition Coefficient Index

PD Proton Density Image

PE Partition Entropy Index

PSO Particle Swarm Optimization

ROI Region Of Interest

SA Simulated Annealing

T1WI T1-Weighted Image

T2WI T2-Weighted Image

SOM Self Organizing Maps

STIR Short Tau Inversion Recovery

TS Tabu Search

WM White Matter

XB Xie-Beni Index

ALGORITMA-ALGORITMA PENGKELOMPOKAN KABUR BERASASKAN CARIAN-HARMONI UNTUK SEGMENTASI IMEJ

ABSTRAK

Algoritma-algoritma pengkelompokan kabur, yang tergolong di dalam kategori pembelajaran mesin tanpa selia, adalah di antara kaedah segmentasi imej yang paling berjaya. Namun demikian, terdapat dua isu utama yang membataskan keberkesanan kaedah ini: kepekaan terhadap pemilihan pusat kelompok permulaan dan ketidakpastian terhadap bilangan kelompok sebenar di dalam set data. Tesis ini bermatlamat untuk menyelesaikan masalah-masalah ini dengan mengunakan algoritma metaheuristik eficien yang dikenali sebagai algoritma Carian Harmoni (HS). Pertama, dua kaedah alternatif pengkelompokan kabur berasaskan HS dicadangkan. Tujuan kedua kaedah ini adalah untuk mengatasi kelemahan algoritma pengkelompokan kabur kovensional yang boleh menghasilkan kelompok suboptimum bergantung kepada pemilihan kelompok permulaan. Kedua, algoritma pengkelompokan kabur berasaskan HS dinamik (DCHS) baharu dicadangkan untuk menganggar bilangan kelompok secara automatik serta memperoleh pembahagian kabur yang baik bagi set data yang digunakan. Kesemua algoritma baharu ini telah diaplikasikan kepada permasalahan segmentasi imej. Pelbagai imej dari domain aplikasi berbeza, termasuk imej dunia sebenar dan sintetik, telah digunakan di dalam tesis ini untuk menunjukkan kebolehgunaan algoritma-algoritma yang dicadangkan. Akhir sekali, DCHS diaplikasikan ke atas dua permasalahan imej perubatan, iaitu segmentasi tumour tulang malignan (osteosarkoma) dan MRI Otak. Hasil kajian menunjukkan kemajuan memberangsangkan berbanding pendekatan lain di dalam domain yang sama.

HARMONY SEARCH-BASED FUZZY CLUSTERING ALGORITHMS FOR IMAGE SEGMENTATION

ABSTRACT

Fuzzy clustering algorithms, which fall under unsupervised machine learning, are among the most successful methods for image segmentation. However, two main issues plague these clustering algorithms: initialization sensitivity of cluster centers and unknown number of actual clusters in the given dataset. This thesis aims to solve these problems using an efficient metaheuristic algorithm, known as the Harmony Search (HS) algorithm. First, two alternative HS-based fuzzy clustering methods are proposed. The aim of these methods is to overcome the limitation faced by conventional fuzzy clustering algorithms, which are known to provide sub-optimal clustering depending on the choice of the initial clusters. Second, a new dynamic HS-based fuzzy clustering algorithm (DCHS) is proposed to automatically estimate the appropriate number of clusters as well as a good fuzzy partitioning of the given dataset. These algorithms have been applied to the problem of image segmentation. Various images from different application domains, including synthetic and real-world images, have been used in this thesis to show the applicability of the proposed algorithms. Finally, the proposed DCHS algorithm is applied to two real-world medical image problems, namely, malignant bone tumour (osteosarcoma) and magnetic resonance imaging brain segmentation. The experimental results are very promising showing significant improvements compared to other approaches in the same domain.

CHAPTER 1

INTRODUCTION

1.1 Data clustering

Data clustering is a technique used to discover patterns and associations within data. It is a multivariate statistical procedure that allows the user to handle large volumes of data and attempts to reorganize these data into relatively homogeneous groups. These groups should allow the user to deal with and utilize the original volume of data more effectively. Accuracy of clustering is essential because it would be counter-productive if the compact form of the data does not accurately represent the original data (Shihab, 2000).

Here, data are a set of points or patterns usually represented as vectors of measurements, features, or points in a multidimensional space. These collected features are combined into a list, which then acts as the input to a chosen computational clustering algorithm. This algorithm then provides a description of the grouping structure that it has discovered within the patterns. The grouping structure, clustering process, is based on some similarity or distance measurements that allocate the given patterns into predefined clusters (classes). Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster.

Data clustering is known as an unsupervised classification technique different from the other classification technique known as supervised classification. The main difference between the two is that in the latter, a collection of labeled (pre-classified) patterns is provided, therefore, the problem is to label a newly encountered, yet unlabeled, pattern. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In other words, no labeled patterns are provided, therefore, the labels associated with clusters are data driven; that is, they are obtained solely from the data (Jain et al., 1999).

Clustering has many useful characteristics, which has made it one of the most popular machine learning algorithms in many domains such as machine learning, artificial intelligence, pattern recognition, web mining, data mining, biology, remote sensing, marketing, image segmentation, etc. (Jain et al., 1999; Brian et al., 2009). During the last several decades, clustering algorithms have proved reliable, especially in categorization tasks that call for semi or full automation (Hore et al., 2008; Zhou and Schaefer, 2009).

1.2 Clustering Approaches

Generally, clustering is a typical unsupervised learning technique used to group similar data points according to some measurement of similarity. This measurement will seek to minimize inter-cluster similarity while maximizing intra-cluster similarity (Jain et al., 1999).

Clustering algorithms can generally be categorized into two groups: hierarchical, and partitional (Jain et al., 1999). The former produces a nested series of partitions, whereas the latter does clustering with one partitioning result. According to Jain et al. (2000) partitional clustering is more popular in pattern recognition applications because it does not suffer from drawbacks such as static-behavior (i.e., data points assigned to a cluster cannot move to another cluster), and the probability of failing to separate overlapping clusters, problems that are prevalent in hierarchical clustering. Partitional clustering can further be divided into two: 1) crisp (or hard) clustering, where each data point belongs to only one cluster, and 2) fuzzy (or soft) clustering, where data points can simultaneously belong to more than one cluster at the same time, based on some fuzzy membership grade. Fuzzy is considered more appropriate than crisp clustering for datasets that exhibit unclear boundaries between clusters or regions (Hore et al., 2008; Kang et al., 2009; Zhou and Schaefer, 2009). The fuzzy clustering approach and its characteristics are discussed later in this thesis.

1.3 Clustering-based Image Segmentation Approach

Image segmentation is the task of subdividing the image into constituent regions, in which each region shares similar feature properties. Image segmentation is thus considered a core process and is one of the most challenging tasks in any computer vision system (Gonzalez and Woods, 2008). Image segmentation plays a major role in many different domains. In medical imaging for instance, image segmentation techniques can assist doctors and radiologists locate tumours and other pathologies, measure tissue volume, diagnose illnesses, aid in computer-guided surgery, treatment planning, surgical simulation, therapy evaluation, and also study anatomical structure. In the pattern recognition domain, image segmentation is used to isolate regions of interest (ROIs) from images containing characters, fingerprints, signatures, faces, and gestures. In remote sensing, regions such as roads, buildings, rivers, etc. can be identified using image segmentation. In the manufacturing industry, those inspecting and assembling manufacturing products also benefit from image segmentation. The foregoing are but few examples from a much larger number of possible applications of image segmentation (Sergios and Konstantinos, 2008).

Such diversities in applications have thus spurred interest from digital image processing experts to develop advanced algorithms to improve segmentation results within a given domain. This is necessary, as different domains not only deal with different types of images, but also demonstrate different image properties. Furthermore, there is the issue of image complexity, which pertains to the amount of subjective information contained in images. Using one algorithm in different areas simply cannot get the job done (Zhang et al., 2008). Thus, many algorithms have been proposed over the last several decades, each of which uses different induction principles (Pal and Pal, 1993; Pham et al., 2000). These algorithms can be categorized into various groups such as thresholding-based, deformable models-based, clustering-based, histograms-based, classification-based, etc. (Pal and Pal, 1993; Pham et al., 2000).

Among these algorithms, fuzzy clustering-based segmentation methods are of considerable benefit because most images exhibit unclear boundaries between their regions. In this context, fuzzy clustering has shown tremendous potential, as it can naturally cope with such data characteristics. It is therefore not surprising that the fuzzy clustering algorithms represented by the fuzzy c-means (FCM) algorithm (Bezdek, 1981), is the most widely used algorithm in numerous image applications (Hore et al., 2008; Kang et al., 2009). Both image segmentation and clustering share the same goal of finding accurate classification of their input. Clustering algorithms consider the image pixels as patterns and each pixel is assigned to a cluster (image region) based on some feature similarity (Rosenfeld and Kak, 1982). In this thesis, the application of the proposed fuzzy clustering algorithms is invested in solving the problem of image segmentation as can be seen in Chapters 5 and 6. In Chapter 7, we describe a specific application of the proposed fuzzy clustering algorithms to two difficult real-world medical image problems.

1.4 Metaheuristic-based Clustering

Metaheuristic algorithms are well-known approximate algorithms that can solve optimization problems with satisfactory results (Blum and Roli, 2003, 2008). They can be defined as " ...high level strategies for exploring search spaces by using different methods " (Blum and Roli, 2003). Metaheuristics thus came forth to overcome the major drawback of well-known approximate algorithms, local search algorithms, that may stop at a very poor quality local optima. Metaheuristic is a general heuristic method applicable to a wide range of different optimization problems. They can be categorized into two classes: local search-based metaheuristic and population-based metaheuristic, where the former is based on evolving a single solution, while the latter is based on a population of solutions. The population-based metaheuristic approach has some advantages over the local search-based metaheuristic approach as seen in Appendix A. The main advantages of population-based metaheuristic algorithms are their abilities to cope with local optima and explore large solution spaces effectively by maintaining, recombining, and comparing several candidate solutions simultaneously. Many are inspired by natural phenomena as in the case of Particle Swarm Optimization (PSO), Simulated Annealing (SA), Genetic Algorithm (GA), and Harmony Search (HS) algorithm. These algorithms are intelligently inspired by natural phenomena to provide efficient solution techniques to yield high-quality solutions in a reasonable time.

HS is a relatively new metaheuristic algorithm developed by Geem et al. (2001) to solve optimization problems. Ever since the emergence of this algorithm, it has been able to attract many researchers to develop HS-based applications in many optimization problems (see (Ingram and Zhang, 2009) and references therein). HS imitates the natural phenomenon of musicians' behaviors when they collectively tune the pitches of their instruments together to achieve a fantastic harmony as measured by aesthetic standards. It is a successful metaheuristic algorithm that can explore the search space of the given dataset in a parallel optimization environment, where each solution (harmony) vector is generated by intelligently exploring and exploiting a search space. It is thus considered a population-based algorithm with local search-based aspects (Geem, 2009b). This feature distinguishes HS, along with other features such as (1) the generation of a new vector after considering all existing vectors, rather than considering only two vectors as in GA (parents); (2) the independent consideration for each decision variable in harmony memory vector; (3) the consideration of continuous decision variable values without any loss of precision; (4) no requirement of decimal-binary conversions or a fixed number (2n) of decision variable values as in GA; and (5) no need for any starting values of the decision variables or require complex derivatives as in gradient-based methods. These make it a preferable technique not only as standalone algorithm but also when combined with other metaheuristic algorithms

Due to the Non-deterministic Polynomial-time hard ,known as NP-hard, nature of partitional clustering methods, where the minimization of an objective function (e.g., the sum-of-squared errors) is the principle of these methods, clustering problems can be classified as optimization problems (Falkenauer, 1998). Therefore, metaheuristic algorithms are widely believed to be used as a clustering algorithm. This is based on the ability of such algorithms to solve NP-hard problems with satisfactory near-optimal solutions and significantly less computational time compared with exact algorithms (Hruschka et al., 2009).

1.5 Mapping between Clustering, Segmentation, and Optimization

For an overview of the relationship between these three domains, image segmentation, clustering, and optimization, it can be said that image segmentation can be considered a clustering problem and the clustering problem can regarded as an optimization problem. The mapping between these domains is described as follows.

Image segmentation-clustering mapping: Simply, each pixel in an image can be mapped as a pattern in the clustering domain, while image regions are also mapped to be clusters or classes. Furthermore, the concept of both domains is the same since their goal is to find the accurate classification of their input.

Clustering-optimization mapping: Both approaches share the same goal of selecting best elements from sets of available alternatives. In optimization, these elements are called decision variables while in clustering they could be called cluster centers (centroid-based), cluster labels (label-based), or cluster medoids (medoid-based). This process of selecting best elements from sets of available alternatives can be achieved by minimizing or maximizing the objective function. Both clustering and optimization share the same goal in terms of minimizing or maximizing of an objective function to reach the best elements.

Generally, in each iteration of the optimization process and based on its evolving process, a new solution vector is generated in which its decision variables are values that represent cluster centers (which represent an image's regions), then a reallocation of each pattern (i.e., pixel) to the nearest region with membership degree (i.e., fuzzy membership $\in [0,1]$) is performed. Once this has been done, a calculation to such solution in terms of objective function is performed, and a judgment on its value makes it accepted or rejected. This process is repeated until the stopping criterion is met. At the end, the optimal clusters with their pixel members represent the segmented image.

After this mapping, a conclusion can be drawn that optimization algorithms such as metaheuristic algorithms are suitable for clustering and image segmentation problems. However, one question remains to be answered namely why an optimization approach, especially metaheuristic, needs to be used. To answer this, it must be admitted that the current approach of clustering algorithms has some weaknesses that plague their performance, and the main solution for such weaknesses, as will be seen in the following section, is to use the metaheuristic approach as a clustering approach.

1.6 Motivation of the Research - Problem Statement

Despite their strengths and popularity as algorithms of choice for clustering purposes, partitional clustering algorithms suffer from some drawbacks. Among the main issues are the following:

1. It is sensitive to the cluster centers initialization step, therefore the tendency to be trapped in local optima is very high.

2. Required prior knowledge of number of clusters for a given dataset.

3. Sensitivity to noise and outliers.

This thesis intends to shed light on two weaknesses present in partitional clustering algorithms, namely 1 and 2, as previously mentioned. However, the proposed algorithms are tested against noise and outliers. A description is provided for both problems as follows:

1.6.1 Cluster Centers Initialization Sensitivity - Local Optima Problem

Selecting the initial cluster centers is considered one of the most challenging tasks in clustering algorithms. Generally, clustering algorithms seek to minimize an objective function, although it is unfortunately guaranteed only to yield local minima (Bezdek et al., 1987; Hathaway and Bezdek, 1986; Selim and Ismail, 1984). Improper selection of initial cluster centers will lead the searching process toward an optimal solution that stays in local optima, and therefore produces an undesirable clustering result. The main cause for this local optimal problem is when search algorithms work in a similar fashion to a hill climbing algorithm (Kanade and Hall, 2007). The hill climbing algorithm is a local search-based algorithm that moves in one direction without performing a wider scan of the search space to minimize (or maximize) the objective function. This behavior prevents the algorithm from exploring other regions in the search space that might have a better, or even the desired solution. Since hill climbing-like algorithms are only guaranteed local optimal solutions, consequently, the same initial cluster centers in a dataset will always generate the same cluster results, and better results might as well be obtained if the algorithm is run with different initial cluster centers.

1.6.2 Prior Knowledge of the Number of Clusters

Most existing partitional clustering techniques (including crisp and fuzzy), are manually supplied with the number of classes (clusters), instead of automatically being determined during execution. In many real-world data, the appropriate number of clusters is normally unknown or even difficult to be approximated subjectively (Hruschka et al., 2009; Maulik, 2009; Das, Abraham, Chakraborty and Konar, 2009; Campello et al., 2009). For example, in automatic medical diagnostic systems (e.g., to detect areas of pathological cells in the brain, bone, or breast), such systems should be able to automatically identify the different types of human cells and tissues (such as valuable cells, necrotic cells, edema cells, normal cells, and other body tissues) in each image. However, the number of tissue types is unknown in each image, since these images might be scanned at different positions of the human body, which may or may not have these types of tissues. Furthermore, these systems normally deal with a huge number of images generated from one or more medical modalities such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). Moreover, different sequences of images are available from these imaging modalities such as T1-Weighted Image (T1WI), T2-Weighted Image (T2WI), Short Tau Inversion Recovery (STIR), Fluid Attenuation Inversion Recovery (FLAIR), etc., where each sequence provides different types of information for the tissues under study. For natural images, the situation may be more difficult since the actual number of regions is normally huge and uncertain. Therefore, depending on the aforementioned, segmentation based on automatic determination of appropriate number of clusters is required and desired. yet it is by no means an easy task to do.

In the spirit of the main cause of the initialization sensitivity problem, the local search behavior, several global-based or improved local-based search algorithms have been proposed in the last few decades to address this problem such as (Kanade and Hall, 2007; Selim and Alsultan, 1991; Al-Sultan and Selim, 1993; Al-Sultan and Fedjki, 1997; Bezdek et al., 1994; Hall et al., 1999; Yong-Guo et al., 2004; Pham et al., 2007; Lili et al., 2007; Mahdavi et al., 2008; Maulik and Saha, 2009). The main advantages of these algorithms are their abilities to cope with local optima and explore large solution spaces effectively by maintaining, recombining, and comparing several candidate solutions simultaneously. These solutions can be categorized into two: 1) using the metaheuristic algorithm to find the appropriate cluster center values, and then using these values as initial values for the clustering algorithms as can be seen in (Hall et al., 1999; Pham et al., 2007; Kanade and Hall, 2007), and 2) using a metaheuristic as a clustering algorithm such as (Mahdavi et al., 2008; Maulik and Saha, 2009). Having thoroughly scrutinized these algorithms and their development, one can say that the competition to improve their performance to solve the clustering problem was the goal behind such development. These advancements, it should be emphasized, were based on improving the natural behavior of the optimization process of these algorithms. For instance, balancing between the exploration and exploitation strategies of metaheuristic algorithms is one way to improve metaheuristic-based clustering algorithms. Using various evolving techniques such as encoding schemes (i.e., real or binary), selection mechanism or crossover and mutation operators is another way of modifying these algorithms to improve their performance and accuracy. Experimenting with new metaheuristics has also contributed to further improvement in the field. However, despite the promising results shown by these algorithms, and since there is no exact solution in such category, the approximation algorithms, it is desirable to develop a new metaheuristic-based algorithm that can improve the performance even further.

For the second clustering problem on prior knowledge of number of clusters, metaheuristic-based clustering algorithms were proposed in the literature under the name of the dynamic clustering approach. Such approach can automatically determine the appropriate number of clusters as well as a good clustering of the given dataset. However, little effort has been made in the direction of this approach during the last several years (Hruschka et al., 2009), as can be seen in the literature review chapter. The same concepts behind such diversity of metaheuristic algorithms are used in this approach as mentioned earlier for the first clustering problem. Despite the promising results shown by these algorithms, it is desirable to develop a new metaheuristic-based dynamic clustering algorithm that can improve the performance even further and explore the ability of the new algorithm to provide satisfactory solutions.

The main motivation for this thesis is to improve the clustering performance and overcome its weaknesses and thus, improve all related clustering-based applications such as image segmentation. Furthermore, exploring the ability of the new metaheuristic, HS algorithm, to solve such problems is investigated. These motivations are summarized in the following section.

1.7 Research Objective

The primary objectives of this thesis can be summarized as follows.

1. To show that HS algorithm can be successfully used to solve difficult problems in image segmentation domain.
2. To develop an HS-based fuzzy partitional clustering algorithm that will overcomes the limitation of the partitional clustering algorithms.
3. To develop an efficient HS-based algorithm that can automatically predict the appropriate number of clusters as well as improve the clustering performance of the fuzzy clustering algorithms.
4. To validate the developed algorithms with benchmarked datasets and then applying them to address real-world medical problems including:

- Automatic MRI brain image segmentation.
- Automatic MRI malignant bone tumour known as osteosarcoma.

1.8 Scope

The scope of the presented work is defined as follows:

1. This study is scoped to fuzzy partitional clustering approach.
2. The application of the proposed clustering algorithms is scoped to the image segmentation problem.
3. The image type used in this thesis is limited to gray scale.

1.9 Overview of Methodology

In this thesis, new fuzzy clustering algorithms based on HS algorithm are proposed. These new proposed algorithms take advantage of the inherited features of HS to avoid local optimal and come up with an appropriate assumption of number of clusters for the test images. The application of these new algorithms to the problem of image segmentation is investigated. Experimental results are then obtained using various synthetic images with well-known characteristics and natural images from different areas such as medical images and remotely sensed satellite images are also used to show the wide applicability of the proposed approaches. The results of state-of-the-art algorithms when applied to the same test images are also reported to show the relative performance of the proposed approaches compared with other well-known approaches. Due to the stochastic nature of the proposed algorithms, all presented results are averages and standard deviations over several simulations.

1.10 Importance of the Study

This study is particularly important as it is closely connected with several major related topics. Such topics, as mentioned earlier, are image analysis and pattern recognition together with their applications. Image segmentation is considered the core of such applications, and the degree of success mainly depends on the segmentation results. Interestingly, in many of these applications, partitional clustering algorithms seem to be the popular choice for image segmentation (Kang et al., 2009; Zhou, 2009; Zhou and Schaefer, 2009; Xiaohe et al., 2008; Zhou and Rajapakse, 2008). However, due to the shortcomings of partitional clustering algorithms, as mentioned in Section 1.6, results are still unsatisfactory.

As a result of these shortcomings, the author has been motivated to investigate ways to improve this particular clustering category for image segmentation. This research thus intends to propose novel clustering algorithms that can overcome the shortcomings of partitional clustering. These algorithms will be beneficial as they can improve the accuracy of segmentation which will,in turn, increase the performance of image analysis and pattern recognition applications.

Medical doctors and radiologists (henceforth referred to as medical experts) can also benefit from this study through the proposed medical systems that can help them automatically delineate ROIs in MRI images. Generally, the quantitative analysis of medical images is challenging, as the segmentation of the structure of interest is the prerequisite to quantification. Manual segmentation of the tissue of interest (e.g., tumour) from each image slice by a trained radiologist, while remaining the most accepted practice, is a laborious and time-consuming process. It is also affected by inter- and intra-observer variations. Automated approaches, on the other hand, are generally considered faster, objective measures and provide accurate tissue quantification and/or tissue classification. In this context, two systems are proposed in this study; one for MRI brain images while the other for osteosarcoma (i.e., malignant bone tumour). In this thesis, the aim is to add a new valuable episode in a chain of improvements directed to this area of research.

1.11 Contributions

The main contributions of this thesis can be summarized as follows.

1. The development of an efficient fuzzy clustering algorithm called Harmony Fuzzy C-Means (HFCM) algorithm consists of two stages. In the first stage, HS explores the search space looking for the near-optimal cluster center values. In the second stage, the output of the first stage is used to initialize the fuzzy clustering algorithm, FCM, where the later performs the clustering. Such algorithm will not suffer from the initialization sensitivity and its local optimal problem, and it is particularly suitable for applications in which there is domain knowledge that can suggest a reasonable value for the number of clusters.

2. The development of another alternative clustering algorithm called Harmony Fuzzy Image Segmentation Algorithm (HFISA) based on using HS algorithm as a clustering algorithm. Each harmony memory vector is actually a fuzzy membership matrix, where its width represents the number of data points in the given dataset and its height represents the number of predefined clusters. Consequently, the evolving process of the HS is directed to find the near-optimal fuzzy membership value for each data point to the predefined number of clusters. This algorithm is an alternative to the HFCM algorithm in which the local optimal problem is avoided and the domain knowledge is available to suggest a reasonable value for the number of clusters.

3. The development of an efficient dynamic clustering algorithm called Dynamic Fuzzy

Clustering using the HS (DCHS). The DCHS algorithm could automatically determine the appropriate number of clusters as well as a good fuzzy partitioning of the given dataset. This algorithm is suitable for applications in which there is no domain knowledge available that can give any indication of what is the proper value of number of clusters.

4. The development of an efficient DCHS-based approach to automatically delineate the ROIs, i.e., white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), of the brain MRI images.

5. The development of an efficient automatic medical image segmentation framework based on the DCHS algorithm to delineate the tumour region of osteosarcoma.

1.12 Thesis Outline

Chapter 2: This chapter provides the basic definitions and concepts of harmony search, image segmentation, and data clustering that will be used in this thesis.

Chapter 3: This chapter provides a review of the state-of-the-art techniques available in the literature for the main topics of this thesis including HS algorithm, image segmentation, and the main weaknesses of the clustering approach.

Chapter 4: This chapter provides an overview of the methodology used to achieve the research objectives of this thesis.

Chapter 5: Presents the first contribution of this thesis, new HS-based fuzzy clustering algorithms. Two alternative methods from different points of view are proposed. The first method consists of two stages. In the first stage, HS explores the search space looking for the near-optimal cluster center values. In the second stage, the output of the first stage is used to initialize the fuzzy clustering algorithm, FCM, where the latter performs the clustering. The second method is based on using HS as a clustering algorithm, in which each harmony memory vector represents a clustering solution. The application of the proposed clustering algorithm to the problem of image segmentation is investigated. To illustrate its wide applicability, it is applied to different types of images such as natural images, synthetic images, medical MR images, and remote sensing images.

Chapter 6: Presents a new dynamic clustering algorithm called DCHS. In this algorithm, the capability of standard HS is modified to automatically evolve the appropriate number of clusters, as well as the locations of cluster centers. This approach is applied to unsupervised image classification and is tested on synthetic, natural, medical, and remote-sensing images. All images were selected to show the wide applicability of the proposed algorithm and to compare its results with state-of-the-art algorithms in the domain of dynamic clustering, such as GA, PSO, and other metaheuristic algorithms.

Chapter 7: This chapter presents two applications of the proposed HS-based algorithms. They are two difficult real-world medical image problems. The first application is the automatic MRI brain image segmentation problem, where the attractive tissues in brain such as WM, GM, and CSF are automatically segmented. The second application is from USM Medical center named as malignant bone tumour (osteosarcoma), in which a new framework based on multi-spectral information from various MRI sequences is proposed. In both applications, a set of experiments was conducted and the results were visually and statistically compared with other state-of-the-art methods.

Chapter 8: Concludes the main components of the proposed work, and presents plans for future research directions.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter presents an overview of three topics namely Harmony search algorithm, image segmentation and clustering. Firstly, an explination of the harmony search algorithm, which is the fundamental element of the work in this thesis, is presented. In the context of the harmony search and clustering, the second part will explain the basic concepts of image segmentation. Thirdly, the theoretical background of data clustering is discussed. Basic definitions of terms will be provided about the topic followed by explanations about clustering techniques with an emphasis on fuzzy clustering within partitional clustering approaches.

2.1 The HS Algorithm

HS (Geem et al., 2001) is a relatively new population-based metaheuristic optimization algorithm that imitates the music improvisation process where the musicians improvise their instruments' pitch by searching for a perfect state of harmony. It has been successfully tailored to various scientific and engineering applications such as music composition (Geem and Choi, 2007), sudoku puzzle solving (Geem, 2007a), tour planning (Geem, Tseng and Park, 2005), web page clustering (Forsati et al., 2008; Mahdavi and Abolhassani, 2009), structural design (Lee and Geem, 2004; Geem, 2009a), water network design (Geem, 2009c), vehicle routing (Geem, Lee and Park, 2005), dam scheduling (Geem, 2007b), ground water modeling (Ayvaz, 2009, 2007), soil stability analysis (Cheng et al., 2008), ecological conservation (Geem and Williams, 2008), energy system dispatch (Vasebi et al., 2007), heat exchanger design (Fesanghary et al., 2009), transportation energy modeling (Ceylan et al., 2008), satellite heat pipe design (Geem and Hwangbo, 2006), medical physics (Panchal, 2009), timetabling (Al-Betar et al., 2008; Al-Betar, Khader and Liao, 2010), RNA structure prediction (Mohsen et al., 2010), etc. For further information on these applications see (Ingram and Zhang, 2009) and references therein. It is a very successful metaheuristic algorithm that can explore the search space of a given data in parallel optimization environment, where each solution (harmony) vector is generated by intelligently exploring and exploiting a search space (Geem, 2009c).

HS as mentioned mimic the improvisation process of musicians' with an intelligent way as can be seen in Figure 2.1. The analogy between improvisation and optimization is likely as follows (Geem, 2010):

1. Each musician corresponds to each decision variable.
2. Musical instrument's pitch range corresponds to the decision variable's value range.
3. Musical harmony at a certain time corresponds to the solution vector at a certain iteration.
4. Audience's aesthetics corresponds to the objective function.

Just like musical harmony is improved time after time, solution vector is improved iteration by iteration. In general, HS has five steps and they are described as in (Geem, Tseng and Park, 2005) as follows:

Step 1. **Initialize the Optimization Problem and HS Parameters**

The optimization problem is defined as follows:

$$\begin{aligned} &\text{minimize/maximize} \quad && f(a), \\ &\text{subject to} && a_i \in \mathbf{A}_i,\ i = 1, 2, \ldots, N \end{aligned} \tag{2.1}$$

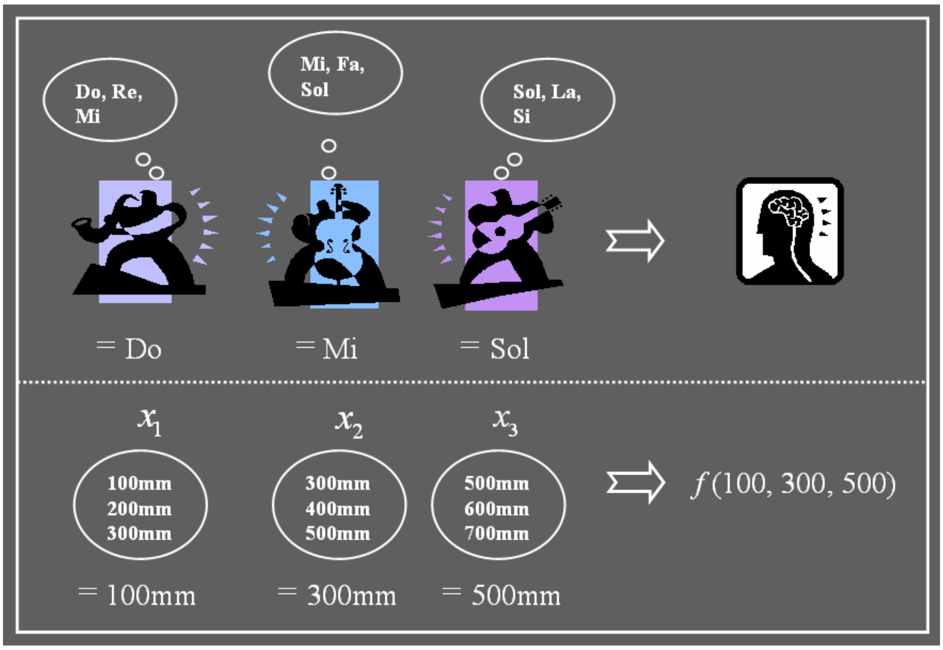

Figure 2.1: Analogy between Improvisation and Optimization, obtained from (Geem, 2010)

where $f(a)$ is an objective function; a is the set of decision variable (a_i); $\mathbf{A}_i$ is the set of possible range of values for each decision variable, ${}_La_i \leq \mathbf{A}_i \leq {}_Ua_i$; ${}_La_i$ and ${}_Ua_i$ are the lower and upper bounds for each decision variable (a_i) respectively. The variable N is the number of decision variables.

Then, the parameters of the HS are initialized. These parameters are:

(a) Harmony Memory Size (HMS) (i.e., number of solution vectors in harmony memory);

(b) Harmony Memory Considering Rate (HMCR), where HMCR $\in [0,1]$;

(c) Pitch Adjusting Rate (PAR), where PAR $\in [0,1]$;

(d) Stopping Criteria (i.e., number of improvisation (NI));

More explanation of these parameters is provided in the next steps.

Step 2. **Initialize Harmony Memory**

The harmony memory (HM) is a matrix of solutions with a size of HMS, where each HM vector represents one solution as can be seen in Eq.2.2. In this step, the solutions are randomly constructed and optionally rearranged in a reversed order to HM, based on their objective function values such as $f(a^1) \leq f(a^2) \cdots \leq f(a^{HMS})$, where $f(a^j)$ is the objective function value for j^{th} harmony memory vector.

$$HM = \left[\begin{array}{cccc|c} a_1^1 & a_2^1 & \dots & a_N^1 & f(a^1) \\ a_1^2 & a_2^2 & \dots & a_N^2 & f(a^2) \\ \vdots & \vdots & \dots & \vdots & \vdots \\ a_1^{HMS} & a_2^{HMS} & \dots & a_N^{HMS} & f(a^{HMS}) \end{array}\right] \tag{2.2}$$

Step 3. **Improvise New Harmony**

This step is the essence of the HS algorithm and the cornerstone that has been laid for building this algorithm. In this step, the HS generates (improvises) a new harmony vector, $a^{'} = (a_1^{'}, a_2^{'}, a_3^{'}, \dots, a_N^{'})$. It is based on three operators: memory consideration; pitch adjustment; or random consideration. In the memory consideration, the values of the new harmony vector are randomly inherited from the historical values stored in HM with a probability of HMCR. Therefore, the value of decision variable $(a_1^{'})$ is chosen from $(a_1^1, a_1^2, a_1^3, \dots, a_1^{HMS})$ that is stored in HM. The next decision variable $(a_2^{'})$ is chosen from $(a_2^1, a_2^2, a_2^3, \dots, a_2^{HMS})$, and the other decision variables, $(a_3^{'}, a_4^{'}, a_5^{'}, \dots, a_N^{'})$, are chosen consecutively in the same manner with the probability of HMCR $\in [0,1]$. The usage of HM is similar to the step where the musician uses his or her memory to generate a good tune. This cumulative step ensures that good harmonies are considered as the elements of new harmony vectors.

Alternativally, where the other decision variable values are not chosen from HM, according to the HMCR probability test, they are randomly chosen according to their possible range, $a_i^{'} \in \mathbf{A}_i$. This case is referred to as random consideration (with a probability of (1-HMCR)), which increases the diversity of the solutions and drives the system further to explore various diverse solutions so that global optimality can be attained.

The following equation summarizes these two steps i.e., memory consideration and random consideration.

$$a_i^{'} \leftarrow \begin{cases} a_i^{'} \in \{a_i^1, a_i^2, a_i^3, \ldots, a_i^{HMS}\} & w.p.\ HMCR \\ a_i^{'} \in A_i & w.p.\ (1 - HMCR) \end{cases} \tag{2.3}$$

For example, a HMCR of 0.95 indicates that the HS algorithm will choose the decision variable value from historically stored values in the HM with a 95% probability or from the entire possible range with a 5% probability. Therefore, if a generated random number $w \in [0,1]$ is equal to 0.78 then, the value of the decision variable is picked out from HM since it is less than the HMCR(0.95).

Furthermore, an additional search for good solutions in the search space is achieved through tuning each decision variable in the new harmony vector, $a^{'} = (a_1^{'}, a_2^{'}, a_3^{'}, \ldots, a_N^{'})$, inherited from HM using PAR operator. These decision variables are examined to be tuned with the probability of PAR $\in [0,1]$ as in Eq.5.3.

$$a_i^{'} \leftarrow \begin{cases} Adjusting\ Pitch & w.p.\ PAR \\ Doing\ Nothing & w.p.\ (1 - PAR) \end{cases} \tag{2.4}$$

If a generated random number $rnd \in [0,1]$ within the probability of PAR then, the new decision variable $(a_i^{'})$ will be adjusted based on the following equation:

$$(a_i^{'}) = (a_i^{'}) \pm rand() * bw \tag{2.5}$$

Here, bw is an arbitrary distance bandwidth, or fret width fw as renamed in the updated version of HS (Geem, 2010), used to improve the performance of HS and $(rand())$ is a function that generates a random number $\in [0,1]$. Actually, *bw* determines the amount of movement or changes that may have occurred to the components of the new vector. The value of *bw* is based on the optimization problem itself i.e., continuous or discrete. In general, the way that the parameter (PAR) modifies the components of the new harmony vector is an analogy to the musicians' behaviors when they slightly change their tone frequencies in order to get better harmonies. Consequently, it explores more solutions in the search space and improves the searching abilities.

All of these operators are well-illustrated using pseudo code as in Fig 2.2.

Step 4. **Update the Harmony Memory**

In order to update HM with the new generated vector $a^{'} = (a_1^{'}, a_2^{'}, a_3^{'}, \ldots, a_N^{'})$, the objective function is calculated for each new harmony vector $f(a^{'})$. If the objective function value for the new vector is better than the worst harmony vector stored in HM, then the worst harmony vector is replaced by the new vector. Otherwise, this new vector is ignored.

$$a^{'} \in HM \wedge a^{worst} \notin HM \tag{2.6}$$

However, for the diversity of harmonies in HM, other harmonies (in terms of least-similarity) can be considered. Also, the maximum number of identical harmonies in HM can be considered in order to prevent premature HM.

Step 5. **Check the Stopping Criterion**

The iteration process in steps 3 and 4 is terminated when the maximum number of improvisations (NI) is reached. Finally, the best HM vector is selected and is considered to be the best solution to the problem under investigation.